Attorney Docket No.: 0160113

UNITED STATES PATENT APPLICATION

FOR

VOICING INDEX CONTROLS FOR CELP SPEECH CODING

INVENTOR:

YANG GAO

CERTIFICATE OF EXPRESS MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service "Express Mail Post Office to addressee" Service under 37 C.F.R. Sec. 1.10 addressed to: Commissioner for Patents, P. O. Box 1450, Alexandria, VA 22313-1450, on 3/11/04

Express Mailing Label No.: EV420421958US

Lori Lapidario — Name

Lori Lapidario — Signature

PREPARED BY:

FARJAMI & FARJAMI LLP
26522 La Alameda Ave., Suite 360
Mission Viejo, California 92691

(949) 282-1000
Customer No. 25700

25700

PATENT TRADEMARK OFFICE

03M0004/US

VOICING INDEX CONTROLS FOR CELP SPEECH CODING

RELATED APPLICATIONS

The present application claims the benefit of United States provisional application serial number 60/455,435, filed March 15, 2003, which is hereby fully incorporated by reference in the present application.

The following co-pending and commonly assigned U.S. patent applications have been filed on the same day as this application, and are incorporated by reference in their entirety:

United States Patent Application Serial Number _____, "SIGNAL DECOMPOSITION OF VOICED SPEECH FOR CELP SPEECH CODING," Attorney Docket Number: 0160112.

United States Patent Application Serial Number _____, "SIMPLE NOISE SUPPRESSION MODEL," Attorney Docket Number: 0160114.

United States Patent Application Serial Number _____, "ADAPTIVE CORRELATION WINDOW FOR OPEN-LOOP PITCH," Attorney Docket Number: 0160115.

United States Patent Application Serial Number _____, "RECOVERING AN ERASED VOICE FRAME WITH TIME WARPING," Attorney Docket Number: 0160116.

BACKGROUND OF THE INVENTION

1. FIELD OF THE INVENTION

The present invention relates generally to speech coding and, more particularly, to Code Excited Linear Prediction (CELP) speech coding.

2. RELATED ART

Generally, a speech signal can be band-limited to about 10 kHz without affecting its perception. However, in telecommunications, the speech signal bandwidth is usually limited much more severely. It is known that the telephone network limits the bandwidth of the speech signal to between 300 Hz to 3400 Hz, which is known as the "narrowband". Such band-limitation results in the characteristic sound of telephone speech. Both the lower limit at 300Hz and the upper limit at 3400 Hz affect the speech quality.

In most digital speech coders, the speech signal is sampled at 8 kHz, resulting in a maximum signal bandwidth of 4 kHz. In practice, however, the signal is usually band-limited to about 3600 Hz at the high-end. At the low-end, the cut-off frequency is usually between 50 Hz and 200 Hz. The narrowband speech signal, which requires a sampling frequency of 8 kb/s, provides a speech quality referred to as toll quality. Although this toll quality is sufficient for telephone communications, for emerging applications such as teleconferencing, multimedia services and high-definition television, an improved quality is necessary.

The communications quality can be improved for such applications by increasing the bandwidth. For example, by increasing the sampling frequency to 16 kHz, a wider bandwidth, ranging from 50 Hz to about 7000 Hz can be accommodated, which is referred to as the "wideband". Extending the lower frequency range to 50 Hz increases naturalness, presence and comfort. At the other end of the spectrum, extending the higher frequency range to 7000 Hz increases intelligibility and makes it easier to differentiate between fricative sounds.

Digitally, speech is synthesized by a well-known approach known as Analysis-By-Synthesis (ABS). Analysis-By-Synthesis is also referred to as closed-loop

approach or waveform-matching approach. It offers relatively better speech coding quality than other approaches for medium to high bit rates. A known ABS approach is the so-called Code Excited Linear Prediction (CELP). In CELP coding, speech is synthesized by using encoded excitation information to excite a linear predictive coding (LPC) filter. The output of the LPC filter is compared against the voiced speech and used to adjust the filter parameters in a closed loop sense until the best parameters based upon the least error is found. One of the facts influencing CELP coding is that voicing degree can significantly vary for different voiced speech segments thus causing an unstable perceptual quality in the speech coding.

The present invention addresses the above analysis-by-synthesis voiced speech issue.

SUMMARY OF THE INVENTION

In accordance with the purpose of the present invention as broadly described herein, there is provided systems and methods for improving quality of synthesized speech by using a voicing index to control the speech coding process.

According to one embodiment of the present invention, a voicing index is used to control and improve ABS type speech coding, which indicates the periodicity degree of the speech signal. The periodicity degree can significantly vary for different voiced speech segments, and this variation causes an unstable perceptual quality in analysis-by-synthesis type speech coding, such as CELP.

The voicing index can be used to improve the quality stability by controlling encoder and/or decoder, for example, in the following areas: (a) fixed-codebook short-term enhancement including the spectrum tilt, (b) perceptual weighting filter, (c) sub-fixed codebook determination, (d) LPC interpolation, (e) fixed-codebook pitch enhancement, (f) post-pitch enhancement, (g) noise injection into the high-frequency band at decoder, (h) LTP Sinc window, (i) signal decomposition, etc. In one embodiment for CELP speech coding, the voicing index may be based on a normalized pitch correlation.

These and other aspects of the present invention will become apparent with further reference to the drawings and specification, which follow. It is intended that all such additional systems, methods, features and advantages be included within this description, be within the scope of the present invention, and be protected by the accompanying claims.

BRIEF DESCRIPTION OF DRAWINGS

Figure 1 is an illustration of the frequency domain characteristics of a sample speech signal.

Figure 2 is an illustration of a voicing index classification available to both the encoder and the decoder.

Figure 3 is an illustration of a basic CELP coding block diagram.

Figure 4 is an illustration of a CELP coding process with an additional adaptive weighting filter for speech enhancement in accordance with an embodiment of the present invention.

Figure 5 is an illustration of a decoder implementation with post filter configuration in accordance with an embodiment of the present invention.

Figure 6 is an illustration of a CELP coding block diagram with several sub-codebooks.

Figure 7A is an illustration of sampling for creation of a Sinc window.

Figure 7B is an illustration of a Sinc window.

DETAILED DESCRIPTION

The present application may be described herein in terms of functional block components and various processing steps. It should be appreciated that such functional blocks may be realized by any number of hardware components and/or software components configured to perform the specified functions. For example, the present application may employ various integrated circuit components, e.g., memory elements, digital signal processing elements, transmitters, receivers, tone detectors, tone generators, logic elements, and the like, which may carry out a variety of functions under the control of one or more microprocessors or other control devices. Further, it should be noted that the present application may employ any number of conventional techniques for data transmission, signaling, signal processing and conditioning, tone generation and detection and the like. Such general techniques that may be known to those skilled in the art are not described in detail herein.

Voicing index is traditionally one of the important indexes sent to the decoder for Harmonic speech coding. The voicing index generally represents the degree of periodicity and/or periodic harmonic band boundary of voiced speech. Voicing index is traditionally not used in CELP coding systems. However, embodiments of the present invention use the voicing index to provide control and improve the quality of synthesized speech in a CELP or other analysis-by-synthesis type coder.

Figure 1 is an illustration of the frequency domain characteristics of a sample speech signal. In this illustration, the spectrum domain in the wideband extends from slightly above 0 Hz to around 7.0 kHz. Although the highest possible frequency in the spectrum ends at 8.0 kHz (i.e. Nyquist folding frequency) for a speech signal sampled at 16 kHz, this illustration shows that the energy is almost zero in the area between 7.0 kHz to 8.0 kHz. It should be apparent to those of skill in the arts that the ranges of

signals used herein are for illustration purposes only and that the principles expressed herein are applicable to other signal bands.

As illustrated in Figure 1, the speech signal is quite harmonic at lower frequencies, but at higher frequencies the speech signal does not remain as harmonic because the probability of having noisy speech signal increases as the frequency increases. For instance, in this illustration the speech signal exhibits traits of becoming noisy at the higher frequencies, e.g., above 5.0 kHz. This noisy signal makes waveform matching at higher frequencies very difficult. Thus, techniques like ABS coding (e.g. CELP) becomes unreliable if high quality speech is desired. For example, in a CELP coder, the synthesizer is designed to match the original speech signal by minimizing the error between the original speech and the synthesized speech. A noisy signal is unpredictable thus making error minimization very difficult.

Given the above problem, embodiments of the present invention use a voicing index which is sent to the decoder, from the encoder, to improve the quality of speech synthesized by an ABS type speech coder, e.g., CELP coder.

The voicing index, which is transmitted by the encoder to the decoder, may represent the periodicity of the voiced speech or the harmonic structure of the signal. In another example embodiment, the voicing index may be represented by three bits thus providing up to eight classes of speech signal. For instance, Figure 2 is an illustration of a voicing index classification available to both the encoder and the decoder. In this illustration, index 0 (i.e. "000") may indicate background noise , index 1 (i.e. "001") may indicate noise-like or unvoiced speech signal, index 2 (i.e. "010") may indicate irregular voiced signal such as voiced signal during onset, and indices 3-7 (i.e. "011" to "111") could each indicate the periodicity of the speech signals. For instance, index 3 ("011") may represent the least periodic signal and

index 7 ("111") may indicate the most periodic signal.

The voicing index information can be transmitted by the encoder as part of each encoded frame. In other words, each frame may include the voicing index bits (e.g. three bits), which indicate the periodicity degree of that particular frame. In one embodiment, the voicing index for CELP may be based on a normalized pitch correlation parameter, Rp, and may be derived from the following equation: 10 log $(1-Rp)^2$, where $-1.0 < Rp < 1.0$.

In one example, the voicing index may be used for fixed codebook short-term enhancement, including the spectrum tilt. Figure 3 is an illustration of a basic CELP coding block diagram. As illustrated, the CELP coding block 300 comprises the Fixed Codebook 301, gain block 302, Pitch filter block 303, and LPC filter 304. CELP coding block 300 further comprises comparison block 306, Weighting Filter block 320, and Mean Squared Error (MSE) computation block 308.

The basic idea behind CELP coding is that Input Speech 307 is compared against the synthesized output 305 to generate error 309, which is the mean squared error. The computation continues in a closed loop sense with selection of a new coding parameters until error 309 is minimal.

On the receiving side, the decoder synthesizes the speech using similar blocks 301-304 (see Figure 5). Thus, the encoder passes information to the decoder as needed to select the proper codebook entry, gain, and filters, ...,etc..

In a CELP speech coding system, when the speech signal is more periodic, the pitch filter (e.g. 303) contribution is heavier than the fixed codebook (e.g. 301) contribution. As a result, an embodiment of the present invention may use the voicing index to place more focus in the high frequency region by implementing an adaptive high pass filter, which is controlled by the value of the voicing index. An architecture

such as the one shown in Figure 4 may be implemented. For instance, Adaptive Filter 310 could be an adaptive filter emphasizing the power in the high frequency region. In the illustration, the weighting filter 420 may also be an adaptive filter for improving the CELP coding process.

On the decoder side, the voicing index may be used to select the appropriate Post Filter 520 parameters. Figure 5 is an illustration of the decoder implementation with post filter configuration. In one or more embodiments, Post Filter 520 may have several configurations saved in a table, which may be selectable using information in the voicing index.

In another example, the voicing index may be used in conjunction with the perceptual weighting filter of CELP. The perceptual weighting filter may be represented by Adaptive filter 420 of Figure 4, for example. As is well known, waveform matching minimizes the error in the most important portion (i.e. the high energy portion) of the speech signal and ignores low energy area by performing a mean squared error minimization. Embodiments of the present invention use an adaptive weighting process to enhance the low energy area. For instance, the voicing index may be used to define the aggressiveness of the weighting filter 420 depending on the periodicity degree of the frame.

In yet another embodiment, as illustrated in Figure 6, the voicing index may be used to determine the sub-fixed codebook. There are possibly several sub-codebooks for the fixed codebook, for example, one sub-codebook 601 with less pulses but higher position resolution, one sub-codebook 602 with more pulses but lower position resolutions, and a noise sub-codebook 603. Therefore, if the voicing index indicates a noisy signal, then the sub-codebook 602 or noisy sub-codebook 603 can be used; if the voicing index does not indicate a noisy signal, then one of the sub-codebooks (e.g.

601 or 602) may be used depending on the degree of periodicity of the given frame. Note that the gain block (codebook) 302 may also be applied individually to each sub-codebook in one or more embodiments.

Further, the voicing index may be used in conjunction with the LPC interpolation. For example, during linear interpolation, the previous LPC is equally important as the current LPC if the location of the interpolated LPC is at the middle between the previous one and the current one. Thus, if the voicing index, for example, indicates that the previous frame was unvoiced and the present frame is voiced, then during the LPC interpolation, the LPC interpolation algorithm may favor the current frame more than the previous

The voicing index may also be used for fixed codebook pitch enhancement. Typically, the previous pitch gain is used to perform pitch enhancement. However, the voicing index provides information relating to the current frame and, thus, could be a better indicator than the previous pitch gain information. The magnitude of the pitch enhancement may be determined based on the voicing index. In other words, the more periodic the frame (based on the voicing index value), the higher the magnitude of the enhancement. For example, the voicing index may be used in conjunction with the U.S. patent application serial No. 09/365,444, filed August 2, 1999, specification of which is incorporated herein by reference, to determine the magnitude of the enhancements in the bi-directional pitch enhancement system defined therein.

As a further example, the voicing index may be used in place of pitch gain for post pitch enhancement. This is advantageous, since, as discussed above, the voicing index may be derived from a normalized pitch correlation value, i.e. Rp, which is typically between 0.0 and 1.0; however, pitch gain may exceed 1.0 and can adversely affect the post pitch enhancement process.

As another example, the voicing index may also be used to determine the amount of noise that should be injected in the high frequency band at the decoder side. This embodiment may be used when the input speech is decomposed into a voiced portion and a noise portion as discussed in pending U.S. patent application serial No. ____, filed concurrently herewith, entitled "SIGNAL DECOMPOSITION OF VOICED SPEECH FOR CELP SPEECH CODING", specification of which is incorporated herein by reference.

The voicing index may also be used to control modification of the Sinc window. The Sinc window is used to generate an adaptive codebook contribution vector, i.e. LTP excitation vector, with fractional pitch lag for CELP coding. In wideband speech coding, it is known that strong harmonics appear in the low frequency area of the band and the noisy signals appear in the high frequency area.

Long-term prediction or LTP produces the harmonics by taking a previous excitation and copying it to a current subframe according to the pitch period. It should be noted that if a pure copy of the previous excitation is made, then the harmonic is replicated all the way to the end spectrum in the frequency domain. However, that would not be an accurate representation of a true voice signal and especially not in wideband speech coding.

In one embodiment, for wideband speech signal when the previous signal is used to represent the current signal, an adaptive low pass filter is applied to the Sinc interpolation window, since there is a high probability of noise in high frequency area.

In CELP coding, the fixed codebook contributes to coding of the noisy or irregular portion of the speech signal, and a pitch adaptive codebook contributes to the voice or regular portion of the speech signal. The adaptive codebook contribution is generated using a Sinc window, which is used due to the fact that the pitch lag can be

fractional. If the pitch lag were an integer, one excitation signal could be copied to the next; however, because the pitch lag is fractional, straight copying of the previous excitation signal would not work. After the Sinc window is modified, the straight copying would not work even for integer pitch lag. In order to generate pitch contribution, several samples are taken, as shown in Figure 7A, which are weighted and then added together, where the weights for the samples is called the Sinc window, which originally has a symmetric shape, as shown in Figure 7B. The shape in practice depends on the fractional portion of the pitch lag and the adaptive lowpass filter applied to the Sinc window. Application of the Sinc window is similar to convolution or filtering, but the Sinc window is a non-causal filter. In the representation shown below, a window signal w(n) is convoluted with the signal s(n) in the time domain, which is an equivalent representation to spectrum of the window W(w) multiplied by the spectrum of the signal S(w) in the frequency domain:

$$U_{ACB}(n) = w(n) * s(n) \leftrightarrow W(w)\, S(w).$$

According to the above representation, low passing of the Sinc window is equivalent to low passing the final adaptive codebook contribution (U_{ACB} (n)) or excitation signal; however, low passing of the Sinc window is advantageous due to the fact that the Sinc window is shorter than the excitation. Thus, it is easier to modify the Sinc window than the excitation; further more, the filtering of the Sinc window can be pre-calculated and memorized.

In one embodiment of the present invention, the voicing index may be used to provide information to control modification of the low pass filter for the Sinc window. For instance, the voicing index may provide information as to whether the harmonic

structure is strong or weak. If the harmonic structure is strong, then a weak low pass filter is applied to the Sinc window, and if the harmonic structure is weak, then a strong low pass filter is applied to the Sinc window.

Although the above embodiments of the present application are described with reference to wideband speech signals, the present invention is equally applicable to narrowband speech signals.

The methods and systems presented above may reside in software, hardware, or firmware on the device, which can be implemented on a microprocessor, digital signal processor, application specific IC, or field programmable gate array ("FPGA"), or any combination thereof, without departing from the spirit of the invention. Furthermore, the present invention may be embodied in other specific forms without departing from its spirit or essential characteristics. The described embodiments are to be considered in all respects only as illustrative and not restrictive.